(To appear in Assessment in Education).

Francis Galton, measurement, psychometrics and social progress

by

Harvey Goldstein

University of Bristol

Abstract

This paper looks at Galton's work from the perspective of its influence on subsequent developments in assessment and especially psychometric latent variable models. It describes how Galton's views on scientific validity and assumptions about data analysis came to be incorporated into later perspectives.

Keywords

Galton, educational measurement, factor analysis, psychometrics

Correspondence

h.goldstein@bristol.ac.uk

Modern cultures are deeply imbued with notions of measurement. Nearly all scientific disciplines depend heavily on mathematical measurement and arguably the social sciences have, relatively, seen the most rapid recent development of quantitative methodology and accompanying measurement regimes. Political debate also involves the use of measurements of all kinds and it is often assumed that the introduction of new measurements or the expansion of existing ones needs little justification. Sometimes, but rather rarely outside of the statistical and other scientific professions, discussion will revolve around the accuracy of measurements and whether they could be refined, but the default assumption is that more measurement is a good thing.

A particularly interesting example can be found in current debates in the UK and elsewhere, around the opening up of Government databases. Thus, in April 2010 a UK Government website (data.gov.uk) made available thousands of official datasets and was warmly welcomed by many 'data freedom' commentators, including an anonymous news item in the Royal Statistical Society journal 'Significance' (June 2010). Without wishing to decry this and similar developments (the author is, after all, a user of such databases) what is intersting is the almost universal assumption that the more data we have, and make public, the better.

Another contemporary example is that of institutional accountability where there has been an enormous increase in performance indicators or 'league tables' ranking the 'performance' of schools, hospitals, universities, individual scientists, police forces etc. This has, of course, been driven partly by the advent of information technology that has made the collection and processing relatively straightforward, but again relies heavily on the assumption that having more data has to be a good thing, even though debate may be needed on what to measure, how to measure it, how to process it and how to display the results.

Yet measurement has not always been so popular, and 100 years after the death of Francis Galton(and also incidentally Alfred Binet) is not a bad time to reflect on how we got here, who helped us to get here and how much of it was worthwhile. In the space of a relatively short paper, little more can be done than to provide a sketch and since it is the Galton centennial, his contribution will be dwelt on. Although the paper will attempt to draw links between Galton's work and subsequent developments, no particular claim is made for any *causal* connections, even though Galton's work certainly did exert a direct influence on many of those who followed him..

One of the motivations driving the work of Galton, and contemporaries such as James Cattell, was the perceived need to identify and nurture those individuals, who by virtue of their potential as scientists, were viewed as essential to the advancement of society. By 'society' they were typically thinking of industrialised Western societies and the principal method of measuring scientific achievement was in terms of the reputation a scientist enjoyed among his (occasionally her) peers. Thus, one of the very first databases to be used for a league table was Cattell's 'American Men of Science' (Cattell, 1906) in which he listed biographical details of 1,000, rising to

34,000 in 1944, American scientists. A major purpose was to use the data to compare 'standards' in different universities, regions etc. In the case of Galton, one of his major concerns was to identify the sources of talented scientists and he became involved in comparing families, schools and nations.

The belief that more and better scientists held the key to prosperity and progress was shared by many intellectuals at the end of the nineteenth century, and not just by scientists themselves. The belief that heredity was the key to this was also a widely held assumption (see, for example, Sutherland, 1984, p33). Galton himself was primarily an empirical experimentalist and his contribution, also taken up by Cattell, was to propose ways in which one could identify such (potential) scientists by virtue of measuring the characteristics of families, institutions, regions, etc. With the development of sophisticated devices for measuring 'mental ability', starting with the Binet-Simon tests in 1905, the identification of talented individuals was about to become the subject of a scientific specialism in its own right – psychometrics.

Collecting data

The design of studies to collect data for making causal inferences about the antecedents of intellectual achievements, was in its infancy at the time Galton began his first serious investigations just after the mid nineteenth century. There were no satisfactory ways of measuring the strength of association between characteristics until Galton effectively invented the correlation coefficient, subsequently placed upon a rigorous mathematical basis by Dickson and developed formally into the more general framework of linear regression by Pearson (Stigler, 1986) – though the term 'regression' itself was invented by Galton to denote the lack of perfect correlation between two measurements.

Galton used his measure to good effect in describing relationships between the characteristics of parents and offspring and between measurements taken on the same individual. In 'Kinship and Correlation' (Galton, 1890) he discusses at length the application of regression to predicting unknown body measurements from known measurements. Galton amassed large data sets which he used to develop his ideas. One of the best known is the data on heights of parents and children (1886a).

Confirming assumptions

Galton interprets the regression phenomenon as a combination of genetic inheritance from a child's parents and genetic inheritance from all his or her ancestors; the former pulling the child's stature towards the parents and the latter towards the population mean (the 'mediocre'). The possibility that the environment exhibits an effect in reducing a (perfect) parental genetic correlation is simply not envisaged. For Galton, the consistency of his (inheritance) theory with his observations was sufficient to justify the theory (but see Wachsmuth et al., 2003 who point out that in fact Galton's data did not actually fit his theory very well – due to an inappropriate use of pooling).

This notion that the consistency of data with a particular statistical model is all important and is one that continues to occupy a central role in applied data analysis. It is present implicitly in Spearman's arguments for the existence of a single intelligence factor (see below) and, as we shall see, surfaces at the end of the 20th Century in some of the arguments for one-dimensional factors underlying test item responses.

Certainly there were others arguing for nurture rather than nature, such as Candolle whom Galton was aware of (Godin, 2007), and Cattell (1906) specifically argued against Galton's interpretation (see below). Nevertheless, any search for alternative explanations did not seem to concern Galton, and certainly would not be encouraged by his strong commitment to eugenics. Galton seemed to be someone who believed that data patterns were reflections of underlying realities that had universal validity. He was extraordinarily impressed, for example, with 'the wonderful form of cosmic order expressed by the "law of error" (referred to in his 1869 book *Hereditary Genius* as the 'law of deviation from an average')…It reigns with serenity in complete self-effacement amidst the wildest confusion'. (Galton, 1886b). Such a statement borders on the mystical.

Galton's delight with the normal distribution (as we now refer to it) persuaded him that it would permeate all manner of measurements and in particular mental measurements. He was so impressed with the way it described stature distributions (in France, Scotland and England) that he insisted it *had* to describe mental measurements also (Galton, 1869, pp 32-33).

> Now, if this be the case with stature, then it will be true as regards every other physical feature —as circumference of head, size of brain, weight of grey matter, number of brain fibres, &c.; and thence, by a step on which no physiologist will hesitate, as regards mental capacity.

Galton then goes on to say that he does not rely upon analogy alone to justify this assertion and quotes the marks from a military examination as being a good approximation to a normal distribution. The data he gives, however, are based upon just over 70 recruits grouped into 8 categories. These are consistent with a normal distribution but hardly strong evidence. Again here we see a use of evidence that does not *contradict* his theory as constituting strong evidence in its favour. All of this remains unchanged in the second, 1892, edition of *Hereditary Genius* and informs his papers espousing eugenics up to his death (see e.g. Galton, 1908).

Another example of Galton's comfort with the idea that non-contradiction of a thesis provides good evidence in its favour is in *Hereditary Genius* (P81, second edition) when he argues that the achievements of the sons of eminent men are far greater than those of the adopted sons of Roman Catholic dignitaries, and that since both have similar advantages by way of nurture, it is the genetic component that must be dominant. He seems quite uninterested in speculating about other explanations. Thus, even leaving aside the possibility that some of the adopted sons might in truth be biological offspring, there is a process of selection made by the dignitaries that could be expected to influence achievement and of course, the environment of an adopted son of a Roman Catholic dignitary is likely to be very different from that of the son of an otherwise eminent man.[1]

Galton was not the only one of his contemporaries to eschew a too detailed search for alternative explanations for their findings. Thus James Cattell, a fierce critic of Galton's emphasis on the dominance of hereditary influences himself, regarded his findings of intellectual excellence in a small number of regions of the United States as

[1] Along with most of his educated contemporaries Galton assumed that eminence was largely a male preserve and that female genes were largely irrelevant.

clear evidence of the overwhelming importance of environmental factors. (Cattell, 1906, P734-735)

> The inequality in the production of scientific men in different parts of the country seems to be a forcible argument against the view of Dr. Galton and Professor Pearson that scientific performance is almost exclusively due to heredity. It is unlikely that there are such differences in family stocks as would lead one part of the country to produce a hundred times as many scientific men as other parts……Differences in stock can scarcely be great enough to account for this; it seems to be due to circumstance.

Like Galton, Cattell seems reluctant to contemplate alternative explanations, for example that the existence of scientific opportunities became concentrated in certain areas as a result of hereditary relationships.

I will argue in a later section that a satisfaction with weak non-contradiction for a theory or assumption recurs in the field of mental measurement since Galton's time, in different guises but fundamentally using the same reasoning. I am reluctant to suggest a straightforward causal influence of Galton on successors, and my thesis does not depend on being able to trace any such causal pathways. Rather, I wish to emphasise the importance of Galton as the first major populariser of these attitudes in mental measurement and then to trace how subsequent generations adopted many of the underlying assumptions, albeit expressed in different terms.

Factor analysis – the next logical step.

A few years before Galton's death, Charles Spearman (1904) published his landmark paper that introduced the factor analysis model. As a result of his empirical studies, building in part on the early experiments of Galton, he concluded that

> all branches of intellectual activity have in common one fundamental function (or group of functions), whereas the remaining or specific elements of the activity seem in every case to be wholly different from that in all the others.

One of these 'fundamental functions' was what he termed general intelligence or 'g' which he regarded as a universal factor underlying mental attributes. It was not long before this claim was contested, most notably by Burt (1909) and then by many others, with Spearman defending his position over the next quarter century. One of the most influential critics was Louis Thurstone (1933) who showed how multiple factors could be fitted to data, each independently representing different 'intelligences'. This work influenced later researchers into 'multiple intelligences, for example Raymond Cattell's 16 personality factors model (Cattell, 1957) and Gardner's (1983) multiple intelligences. We shall return to this issue below.

Nevertheless, the notion of a single intelligence proved to be very attractive. A particularly influential promoter of this notion, in the form of the intelligence quotient (IQ), was Lewis Terman. He was instrumental in the development and use of group intelligence tests, the best known of which is the 'Stanford-Binet' based upon the original Binet-Simon tests. He used this on US army recruits in the first world war and believed, as did Galton, that 'intelligence' was an inherited characteristic; essentially one-dimensional. From this assumption it was a small step to argue, as Terman did, that intelligence tests could be used to categorise children and ethnic groups, as well as army recruits, in order to assign them to suitable roles in society. In

the UK this provided intellectual support to the use of IQ tests for educational selection.

In the factor model each of a set of measured or observed variables is assumed to be linearly related to one (or more) unobserved or 'latent' variables or factors (see for example equation (2) below). Such factors may then be interpreted in various ways, most commonly as being characteristics of individuals. Applied to tests of cognitive functioning such models enabled psychologists to build a mathematical foundation for mental measurement. This enabled Galton's ideas to be realised in practice, even down to the assumption that the factors were assumed to have normal distributions – an assumption that remains paramount to this day.

For Spearman, and for his successors for at least the next half century the factor analysis model was based on the assumption that the observed variables themselves had (multivariate) normal distributions and that the relationships being studied were linear. Subsequent developments in item response modelling (Lord and Novick, 1968) extended the range of these models, essentially by allowing discrete measurements such as correct/incorrect variables to be used within a generalised linear model framework (Goldstein and Wood, 1989). Similarly, starting with Thurstone (1933), the factor model itself was generalised to incorporate relationships among multiple factors as well as 'regression' adjustments for other measured factors – what Spearman (see below) referred to as 'irrelevant factors'. The landmark paper here is that of Joreskog (1970) and subsequently these models have been developed and incorporated into computer packages under the general heading of structural equation modelling (SEM) (See e.g. Muthen and Muthen, 2002).

Adjusting for irrelevant factors

One of the interesting features noted by Spearman was the need to discount 'irrelevant factors'. He states

> individual circumstances as after birth materially modify the investigated function are irrelevant and must be adequately eliminated (P 227).

He suggests that age, gender and experience are the relevant factors. He discusses the use of partial correlations to do this, but does not go beyond these factors to consider wider social variables such as income, material deprivation, etc. Nor do such concerns seem to enter the debates around dimensionality – whether a single 'g' or multiple factors such as suggested by Thurstone (1933) and later by Gardner (1983). This mirrors Galton's own preference for choosing to regard conformity of data with a particular model as sufficient justification for acceptance of the model. Such unconcern with attempting to explain observed correlational patterns may be partly explicable by the nature of psychometrics as a discipline, but it does imply a somewhat artificial context for these debates. Interestingly, the term used in many contemporary factor analyses 'confirmatory factor model', seems to be an explicit recognition of this viewpoint. It is also something that is rather foreign to contemporary practice in social science.

It is perfectly possible for high intercorrelations among cognitive abilities to be introduced, at least in part, as a result of variation in 'confounding' variables, say, material or nutritional circumstances (as well as age and gender) that have a common effect on mental functioning.The following artificial example illustrates how this can occur.

Suppose we have a set of 5 test scores and the *j*-th score is generated as follows:

$$y_j = x + e_j$$

$$x \sim N(0,1) \quad e_j \sim N\left(0, \sigma_j^2\right), \qquad \sigma_j^2 = \frac{0.5}{j} \tag{1}$$

where the random variables are mutually independent. The variable *x* represents the 'confounding' variable.

This yields the following population correlation matrix

Table 1

1				
0.73	1			
0.76	0.83	1		
0.77	0.84	0.87	1	
0.78	0.85	0.88	0.90	1

This matrix is consistent with a single common factor as we would expect, for example by studying the tetrad differences. We now randomly generate 1000 sets of 5 responses from model (1) and fit a simple common factor model (using MCMC with diffuse priors) that omits *x*

$$y_{ij} = \mu_j + \lambda_j \theta_i + e_{ij} \tag{2}$$

The results are given in the first column of Table 2, omitting the estimates for the residual variances.

Table 2

Parameter	Model (2)	Model (3)
μ_1	0.01	0.01
μ_2	-0.01	0.00
μ_3	-0.01	0.00
μ_4	-0.01	0.02
μ_5	0.00	0.01
β_1	-	1.03
β_2	-	0.97
β_3	-	1.00
β_4	-	1.02
β_5	-	0.99

λ_1	-1.03	0.02
λ_2	-0.98	-0.06
λ_3	-1.00	0.00
λ_4	-1.02	-0.01
λ_5	-1.00	-0.01

We also fit model (3) where we now adjust for the 'confounding' factor *x*, for example deprivation

$$y_{ij} = \mu_j + \beta_j x_i + \lambda_j \theta_i + e_{ij} \qquad (3)$$

With the results in column 3 of table 2. As is clear from the loadings the common factor now disappears, having been explained by *x* and we recover essentially the model we started with. Thus, while this is not necessarily a very realistic scenario in practice, it does illustrate how a failure to adjust for potential confounders might be misleading.

This problem is certainly understood within the factor analysis literature, but most often is associated with procedures for establishing 'invariant' factor structures in different populations, rather than attempting to account for a factor structure in terms of the variables defining those populations (see for example, Millsap and Meredith, 2007).

Although, somewhat tangential to my argument, it is worth noting that Godfrey Thomson argued against Spearman's inferences by pointing out that the correlational structures observed and claimed by Spearman to demonstrate the existence of a single factor, could also be obtained by a quite different set of assumptions – his theory of 'bonds' (see Bartholomew, 2009). Thomson was not, however, concerned with confounding factors as alternative explanations.

Contentment with establishing merely that a data set is consistent with a particular model is mirrored in the literature by the widespread use of 'goodness of fit' tests to demonstrate such conformity. I am not referring to the inappropriate use of such tests when the sample size is small so that there is very little power to reject such a test, but rather to a general failure to search for confounders even when data are extensive. This lack of concern with wider explanations is a particular feature of item response models, or 'item response theory' as its proponents insist on labelling it. In the 'classic' text in this area Lord (1980) has nothing to say on this issue apart from a very brief reference to comparisons of item response curves between populations. The specific literature on the 'Rasch' model, a particularly simple item response model, is not only content with ignoring confounders but also insistent that only a single dimension is needed in any given application, and a general unwillingness to explore further (See Goldstein, 1980 for an illustrative example). Indeed, proponents of this model regard the model as paramount and suggest that data should be constructed or modified to satisfy the model's assumptions. Thus, Andrich (2004) claims that this model satisfies the conditions of 'fundamental measurement' and as such attains the status of measurement in the physical sciences – a view about measurement in the social sciences that in a slightly different context Gould (1981) has labelled 'physics envy'.

Validity

In many ways psychometric test theory is a curious discipline. One particularly intriguing feature is the way in which any newly devised test is justified in terms of its 'validity'. There are, of course, many facets to the term 'validity' but one of the established methods for ascertaining validity is to correlate the results of a new test with an existing test that that it aims to replace, for example on the grounds of greater relevance. Thus, for example, a new 'cognitive ability' test would be judged partly on how highly it correlated with existing measures of cognitive ability. The same would be true for many educational tests and those measuring general or specific intelligences. It is clear that there is a practical element to this in that a new intelligence test that had a very low correlation with existing established tests would be unlikely to prosper. Moreover, it is not just the existence of a high correlation but also that particular group differences are maintained, for example between males and females for spatial ability tests. It is not difficult to see that such requirements will tend to impose a kind of historical determinism. Thus, if the very first test of spatial ability to be accepted found a particular difference between males and females, whether because of the way items were worded, or for example whether because of the particular samples used to standardise it, every following test would tend to find a similar difference simply because of the requirements for establishing its validity.

From time to time tests have been constructed, using careful selections of items, to show that group differences can be reversed. A well-known example is due to Williams (1972) who constructed an intelligence test in which Black students outperformed White students in contradiction to the usual pattern. Likewise, in the early days of intelligence testing in English schools, it was apparent to the test constructors that they could construct tests to show males outperforming females or vice versa, but in fact chose to confirm existing 'knowledge'. Going back to Galton's era, especially with the particular views on the relative abilities of males and females that were generally accepted at that time and incorporated into test construction, we might well suppose that the historical determinism built into test construction methods has perpetuated many of those views in the results obtained from contemporary tests. It would be an interesting research project to trace the evolution of particular tests from such a standpoint.

More measurement the better?

The rapid development in information technology over the last 30 years has generated enormous possibilities for the collection, processing and publication of social data of all kinds. In countries such as Denmark there exist linked files for all citizens containing personal data about education, health, employment etc. In the UK there is a national pupil database which tracks moves and achievements of every pupil in the maintained education system. Data on crime, health outcomes and educational test scores are published in the UK in the form of league tables where individual schools or hospital units or police forces are regularly ranked.

There is no doubt that the existence of such data, where it has been collected with care and reliably, provides a valuable resource for policymaking, for research and for informing citizens about society. Indeed, there is a continuing debate about how to ensure reliability and relevance and how to avoid misleading inferences when such data are published (British Academy, 2011). Yet almost all of this debate happens among professionals and there is little public discussion of these issues, through the

media. Consider, for example, the issue of 'value added' and 'uncertainty intervals' associated with school league tables. These rarely are a matter for any kind of wide public debate, despite their crucial role in interpreting such league tables. Policymakers in general tend not to see that they have any responsibility to 'educate' the public and highly visible public advocates of 'data freedom' such as Heather Brooke (2010) fail to touch on such issues.

We don't of course know how Galton would react in current circumstances. What we do know, however, is that he was obsessed with the collection of data. He attempted, for example, to measure the regional distribution of 'beauty' by noting how many pretty women he encountered on his travels – another early example of a league table with London at the top and Aberdeen at the bottom(Galton, 1908). With this, as with other measurements, he appealed to what he perceived was the self-evident nature of the validity of the measurement process. This is clear from his hereditary studies, but also in the case of his 'beauty map' he remarks

> Of course this was a purely individual estimate, but it was consistent, judging from the conformity of different attempts in the same population.(P 316).

To modern scientific eyes, this apparent lack of self-questioning and concern for alternative explanations or viewpoints, seems strange and naïve. Indeed, they would presumably have seemed so to many practising scientists of Galton's time, not least to his cousin Charles Darwin. Yet in modern *public* discourse, especially as filtered through most of the popular media, such attitudes, I would maintain, are the norm. Measurement itself, especially if carried out using sophisticated instruments or analysed using complex methodology, is. seen to have the attributes of 'science', and often taken effectively as a justification for believing the results that are presented as if they have a meaningful relation to whatever social process they are claimed to measure.

Conclusions

Using Galton and his views as a starting point I have attempted to describe some of the ways in which psychometric test theory and attitudes towards measurement have tended to develop over the last 100 years. From the outset, with the work of Spearman and others, one of the principal concerns of psychometrics was to be 'scientific'. It sought to do this by developing data collection methods, elaborating its mathematical models and, perhaps most importantly attempting to devise claims or theories that could be 'falsified'. Thus, Horn and McArdle (2004) argue that Spearman's theory was scientific because it was falsifiable in the sense that it would be possible in principle to find data that did not conform to a single 'g' factor. The possibility of falsifiable data, however, is restricted to the particular factor model being used, and as I have argued, very little attempt seems ever to have been made at falsification using confounding variables. In other words, the scientific case is founded on the assumption that the model has to be of the form (2) rather than (3). From the perspective adopted in this paper the scientific status of mental measurement seems more often asserted than real.

I have emphasised the way in which Galton, and many of those who came after, restricted criticisms of their 'models' within a set of ideas that excluded wider social considerations. Interestingly, this aspect of the history of psychometrics seems to have

been little noticed or commented upon by psychometricians themselves, most of whom still appear to be working within the parameters set out around the time of Galton; typically either developing more complex modelling and analysis procedures or devising new measuring instruments. Indeed, a large number of well-known and influential psychometricians, including Raymond Cattell, John Carroll, Hans Eysenck, Arthur Jensen, Robert Lynn and Robert Thorndike, as recently as 1994 published an article in the *Wall Street Journal* (December 13, 1994) defending Herrnstein and Murray's book *The Bell Curve* (Herrnstein and Murray, 1994). In the article, in what seems a throwback to the early 20th Century, they claimed that "Intelligence tests are not culturally biased" and "Intelligence is a very general mental capacity".

What of the future? In some ways there are good grounds for pessimism. Much of the development of mental testing since Galton's time has become fixed, either through the educational texts that are current or, perhaps more importantly through the very large commercial interests associated with the testing industry, who, by and large, do not seek innovation unless it is likely to bring financial reward (see Goldstein, 1989 for an extreme example involving Educational Testing Services and the Golden Rule insurance company). On the other hand my impression is that psychometrics, while it still has a core set of practitioners with their own journals, has moved some way to embrace both the wider statistical community and the wider social science community.

In the area of data collection and presentation at the present time, likewise, I see little ground for optimism. Even in those societies, such as parts of Australia, where crude league tables used to be eschewed, increasing political and commercial pressures seem to be gaining the upper hand. New technologies such as powerful dynamic computer graphics do have the potential to convey findings and patterns in powerful ways, but whether they are used to inform rather than merely impress, remains an open question.

Perhaps the most that one can hope for is that we could reflect more on Galton and his legacy. In particular, a better understanding is needed of the difference between data that 'confirms' a theory by providing a good model fit, and data that allows us to explain observed data patterns using as much potentially falsifiable information as possible.

Acknowledgements

I am extremely grateful to John Bynner and to Gillian Sutherland for comments on an early draft.

References

Andrich, D. (2004). "Controversy and the Rasch model." *Medical Care* **42**: 1-7.

Bartholomew, D., , Deary, I. and Lawn, M. (2009). "Sir Godfrey Thomson: a statistical pioneer." *J Royal Statistical Society*, A**172**: 467-482.

British Academy, (2011). *League tables and performance indicators in the public sector*. London, British Academy.

Brooke, H. (2010). *The Silent State*. London, Heinemann.

Burt, C. (1909). Experimental tests of general intelligence. *British Journal of Psychology, 3*, 94-177.

Cattell, J. M. (1906). "A Statistical Study of American Men of Science: the Selection of a Group of one Thousand Men." *Science* **24**(621): 658-665.

Cattell, R.B. (1957). Personality and motivation structure and measurement. New York, World Book.

Galton, F. (1869). *Hereditary Genius: An inquiry into its laws and consequences*. London, Macmillan.

Galton, F. (1886a). "Regression towards mediocrity in hereditary stature." *Journal of the Anthropological Institute* **15**: 246-263.

Galton, F. (1886b). "Presidential address." Journal of the Anthropological Institute **15**: 489-499.

Galton, F. (1890). "Kinship and Correlation." *North American Review* **150**: 419-431.

Galton, F. (1908). *Memories of my Life*. London, Methuen.

Gardner, Howard (1983) *Frames of Mind: The theory of multiple intelligences,* New York: Basic Books.

Godin, B. (2007). "From Eugenics to Scientometrics: Galton, Cattell and men of Science." *Social Studies of Science* **37**: 691-728.

Goldstein H (1989), 'Equity in Testing After Golden Rule'. Paper read to American Educational Research Association meeting San Francisco, March 27-31.

Goldstein, H. and R. Wood (1989). "Five decades of item response modelling." *British Journal of Mathematical and Statistical Psychology* **42**: 139-167.

Gould, S. J. (1981). *The Mismeasure of Man*. New York, W. W. Norton.

Herrnstein, R. and C. Murray (1994). *The Bell Curve: Intelligence and class structure in American Life*. New York, Free Press.

Horn, J. L. and J. J. McArdle (2007). "Understanding Human intelligence since Spearman.in *Factor analysis at 100*. R. Cudeck and R. C. MacCallum. London, Lawrence Erlbaum.

Joreskog, K. (1969). "A general Approach to confirmatory Maximum Likelihood Factor Analysis." *Psychometrika* **34**: 51-75.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey, Lawrence Erlbaum Associates.

Lord, F. M. and M. R. Novick (1968). *Statistical theories of mental test scores*. Reading, Massachusetts, Addison-Wesley Publishing.

Mackenzie, D. (1981). *Statistics in Britain 1865-1930*. Edinburgh, Edinburgh University Press.

Millsapp, R. E. and W. Meredith (2007). “Factorial invariance: Historical perspectives and new problems.” in *Factor analysis at 100*. R. Cudeck and R. C. MacCullum. Mahwah, New Jersey, Lawrence Erlbaum Associates**:** 131-152.

Muthen, B. O. (2002). "Beyond SEM: General latent variable modelling." *Behaviormetrika* **29:**: 81-117.

Spearman, C. (1904). "'General intelligence' objectively determined and measured." *American Journal of Psychology* **15**: 201-93.

Stigler, S. M. (1986). *The History of Statistics*. Cambridge, Mass, Harvard University Press.

Sutherland, G. (1984). Ability, Merit and Measurement: Mental testing and English Education 1880-1940. Oxford, Clarendon Press.

Williams, Robert L (September 1972). "The BITCH-100: A Culture-Specific Test". American Psychological Association Annual Convention. Honolulu, Hawaii.

Thurstone, L.L. (1933). The vectors of the mind, Psychological review, 41, 1-32.